水环境模型与大数据技术融合研究

马金锋[1]，饶凯锋[1]，李若男[1,2]，张京[1]，郑华[1,2]
1. 中国科学院生态环境研究中心城市与区域生态国家重点实验室，北京 100085；
2. 中国科学院大学，北京 100049

摘要
水环境模型内部结构复杂且计算耗时，造成参数率定、多情景分析及决策优化过程中面临高负荷计算难题，这极大地限制了其应用价值的发挥。如何融合水环境模型和大数据技术，深入挖掘模型应用潜力和充分发挥其应用价值是一个研究热点。总结了水环境模型在实际应用过程中面临的瓶颈，分析了大数据技术在解决这些问题上具有的潜力。基于现有成熟的大数据技术，提出了水环境模型与大数据技术融合框架，解决了水环境模型规模计算、规模存储和应用分析问题。阐述了模型与大数据技术融合过程中面临的问题，提出了具体的实现技术思路。通过SWAT模型率定应用案例，证明融合框架的可行性。最后探讨了大数据背景下水环境模型的未来研究方向，指出开展复杂水环境模型的代理模型研究和水环境模拟优化框架研究是未来的发展趋势。

关键词
水环境模拟；大数据；Hadoop；MapReduce；融合

中图分类号：TP311　　文献标识码：A　　doi: 10.11959/j.issn.2096-0271.2021064

Research on the integration of water environment model and big data technology

MA Jinfeng[1], RAO Kaifeng[1], LI Ruonan[1,2], ZHANG Jing[1], ZHENG Hua[1,2]
1. State Key Laboratory of Urban and Regional Ecology, Research Center for Eco-Environmental Sciences, Chinese Academy of Sciences, Beijing 100085, China
2. niversity of Chinese Academy of Sciences, Beijing 100049, China

Abstract
Applications of water environment models are greatly limited by complex internal structure of the model and time-consuming calculations, significant computation burdens arise during the process of parameter calibration, multi-scenario analysis, and decision-making optimization. How to integrate water environment model and big data technology, deeply explore the potential of model application and give full play to its application value is a research hotspot. The bottlenecks

faced by the water environment model in the process of practical application were summarized, and the potential of big data technology in solving these problems was analyzed. Based on the existing big data technology, a framework for the integration of water environment model and big data technology was proposed to solve the problem of large-scale calculation, large-scale storage and application analysis of water environment model. The problems faced in the integration of model and big data technologies were described, and specific technical ways of implementation were proposed. A case study for calibration of SWAT model was used to demonstrate feasibility of the proposed framework. Finally, the future research direction of water environment modeling in the context of big data was discussed, and the conclusion was pointed out that the research on surrogate modeling of complex water environment model and on water environment simulation and optimization framework is the future development trend.

Key words
water environment simulation, big data, Hadoop, MapReduce, integration

1 引言

模型是集成和综合不同观测数据、理解复杂的交互作用和测试假设，以及模拟历史、预测未来系统发展轨迹和决策如何应对未来趋势的重要工具[1]。根据产生的来源，模型大体可被分为数据驱动和模型驱动两类，数据驱动模型（机理模型）基于关联关系构建，模型驱动模型（机理模型）基于因果关系构建。数据驱动模型是大数据价值体现链条中的重要环节，大数据的核心价值在于寻求或构建合适的模型，利用模型表达事物内在变化规律的过程。在大数据的原生定义中，基于事物之间的关联关系寻求和构建模型，模型构建的成败十分依赖数据的数量和质量。此外，由于数据驱动模型是基于关联关系构建的，其模拟结果无法给予合理解释，导致其认可度不高，因此数据驱动模型通常也被称为“黑箱模型”[2]。目前在水环境领域中，由于可用数据数量少、数据质量低等原因，基于大数据技术成功构建的数据驱动模型案例并不多，总体上处于探索和发展阶段。相比数据驱动模型，水环境领域中的机理模型相对成熟和完善，得到广泛的推广和应用。然而，在大数据环境下，如何从新的视角审视已成熟的机理模型，探索其在大数据技术背景下的价值发挥是一个值得探讨的热点问题。

顾名思义，机理模型从因果关系出发寻找规律，是真实水环境系统的抽象和概化。水环境机理模型是对水体中污染物随空间和时间迁移的转化规律的描述，是一个描述物质在水环境中的混合、迁移过程的数学方程，即描述水体中污染物与时间、空间的定量关系[3]。基于微分方程的水环境机理模型在过去的数十年间取得了极大发展，已经成为水资源及环境管理决策的有力工具。相对于数据驱动模型而言，机理模型除了具备模拟结果可解释、广泛认同和成熟应用的特点，还可以通过开源或者商业的方式获取，即模型的可获得性，这是机理模型区别于数据驱动模型的一个明显特点。数据驱动模型需要耗费大量计算资源来训练和构建，其核心在于如何创建模型；机理模型经过几十年的发展，已相对成熟和完善，其核心在于如何应用模型。相对于数据驱动模型而言，机理模型的可获得性、模拟结果的可解释性、科学界广泛的认同和实际中已有的成熟应用等特点共同决定了深度挖掘机理模型的应用潜力和充分发挥其应用价值是未来研究的重点方向。

在实际应用过程中，机理模型普遍面临

大规模情景运算、模拟结果海量存储和高效分析的难题，这极大地限制了模型的推广和应用，因此，迫切需要探索新的技术和方法来解决这些难题。大数据技术在解决上述难题方面具有潜在优势，研究水环境模型融合大数据技术能否解决和如何解决上述难题是目前面临的一个挑战。本文以水环境模型为例，分析了该模型在实际应用中面临的瓶颈；针对这些瓶颈，分别从规模计算、规模存储和应用分析3个角度，提出了大数据技术与机理模型融合的技术思路，阐述了水环境模型与大数据技术融合的实现流程，以SWAT（soil and water assessment tool）模型率定为应用案例证明了框架的可行性；最后讨论了水环境模型在大数据背景下未来的研究方向。

2 水环境模型应用过程中面临的瓶颈

众所周知，基础数据难以获取以及模型率定、模型验证和场景分析中的高负荷计算是限制模型成功应用的主要瓶颈，如图1所示。水环境模型构建要求有足够的基础数据用于建模、校准和验证。基础数据（如地形、风速、外部污染负荷、流入、流出和开边界条件等）主要作为模型输入，也可为校准模型参数提供依据，评估模型是否能充分描述水体特点。模型需要的数据应尽量准确，数据的局限性会限制模型的应用，数据的质量和数量在很大程度上决定了模型应用的质量。实际上，能够获取的数据往往很少，精准的长期监测是解决数据匮乏的主要途径。此外，理论和经验方法也经常用于弥补数据的欠缺[4]。

模型的核心价值在于对现实世界历史的重现、对未来的预测和对未来优化决策的响应。模型的率定反映了对历史的还原能力。由于水环境数值模型是对真实水环境系统的抽象和概化，模型的参数、输入数据和模型结构均存在不确定性。为了更加客观地反映自然水体中的一系列生化、生物反应过程，基于机理的数值模型在开发过程中不可避免地会引入大量参数。受

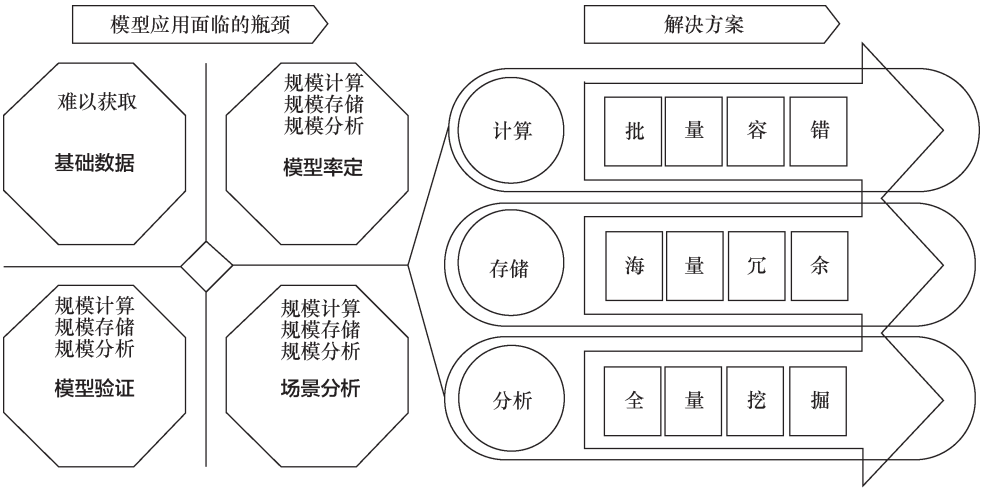

图1 水环境模型应用过程中面临的瓶颈及潜在解决方案

监测资料和对复杂生态过程认知的限制，模型参数的率定往往存在较大困难，使得模型率定成为一个长期的研究方向[5-6]。与此同时，大量的应用不断促进水环境模型的发展，模型变得日益复杂，需要考虑和包含更多的反应过程，增大了模型率定的难度。模型率定是一个严重依赖高性能计算的迭代过程，不同参数组合需要执行不同的独立计算。为了对所有参数组合场景进行统一分析处理，需要对所有独立计算的结果进行统一存储和分析。参数率定、模型验证以及情景分析等都依赖于大规模计算的支持[7]。在目前的实际应用过程中，由于计算规模大，单机多处理器模式和集群并行系统[8]并不能满足上述需求，因此需要探索新的应用模式。大数据技术在支撑规模运算、海量存储和高效分析方面具有显著优势，有望解决上述模型应用中面临的困境。

3 水环境模型与大数据技术融合框架

水环境模型与大数据技术融合体现在分布式计算、存储和分析3个方面，如**图2**所示。针对分布式计算，机理模型与大数据融合体现在模型如何适应分布式并行计算以实现高性能计算。谷歌公司在2004年公开的MapReduce分布式并行计算技术是新型分布式计算技术的代表。典型的MapReduce系统由廉价的通用服务器构成，通过添加服务器节点可线性扩展系统的总处理能力，在成本和可扩展性上都有巨大的优势。造成大数据挖掘革命的技术之一是Hadoop平台上的MapReduce编程模型，其用于在对硬件要求不太高的通用硬件计算机上构建大型集群，从而运行应用程序[9]。除了MapReduce，还有其他分布式计算框架，比如内存迭代

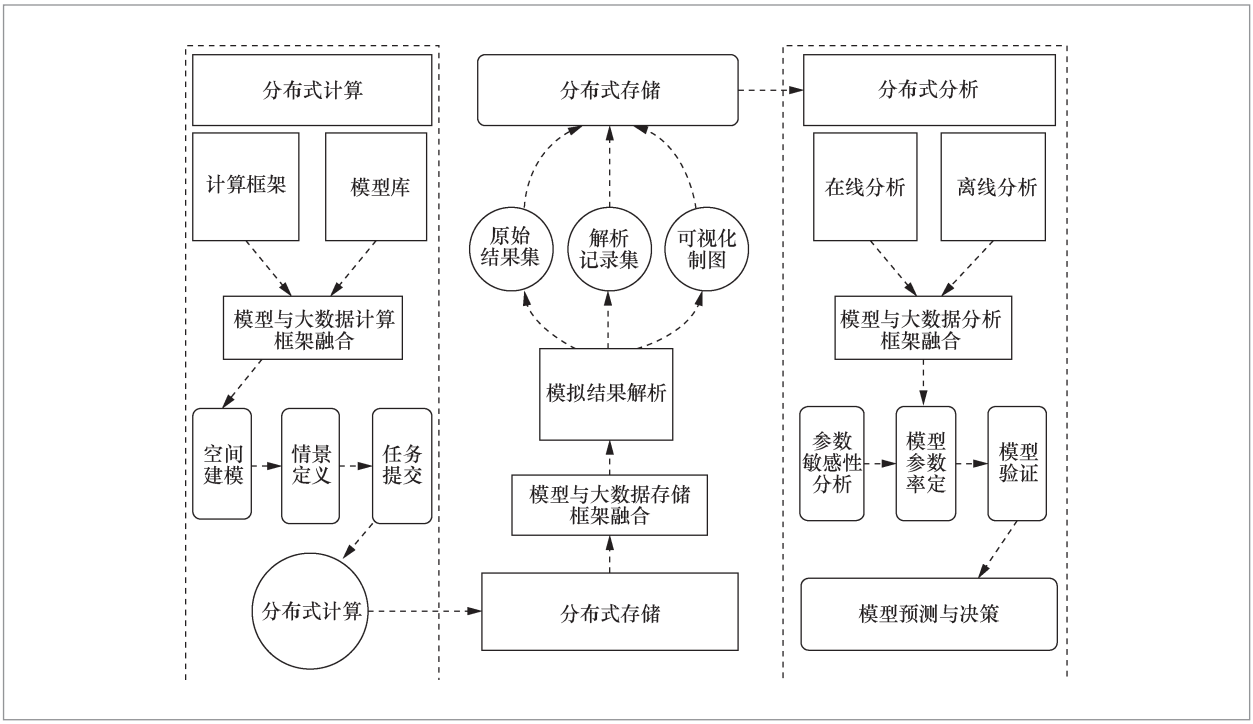

图 2 水环境模型与大数据技术融合框架

计算框架Spark和流式计算框架Storm等。MapReduce属于离线式批量计算框架，鉴于数值模型具有CPU密集型计算的特点，该模型适合采用MapReduce框架。对于计算结果的交互式查询分析，则适合采用Spark框架[8]。大数据计算框架与机理模型融合的核心在于将批量模型算例文件分发到计算节点，模型计算程序定位算例文件所在节点，启动计算程序执行计算。

针对模型模拟结果海量存储，机理模型与大数据融合体现在模型结果（包括原始结果和解析结果）如何实现高效持久化存储。在存储方面，2006年谷歌提出的文件系统GFS以及随后的Hadoop分布式文件系统（Hadoop distributed file system，HDFS）奠定了大数据存储技术的基础。与传统存储系统相比，GFS和HDFS将计算和存储节点在物理上结合在一起，从而避免在数据密集计算中易形成的I/O吞吐量的制约。同时这类分布式存储系统的文件系统也采用了分布式架构，可以达到较高的并发访问能力。GFS和HDFS属于底层的文件存储模式，为了支持非结构化数据存储，BigTable和HBase诞生了。其中，HBase是一个针对结构化数据的可伸缩、高可靠、高性能、分布式和面向列的动态模式数据库。和传统关系数据库不同，HBase采用BigTable的数据模型，即增强的稀疏排序映射表（key/value），其中，键由行关键字、列关键字和时间戳构成。HBase提供了对大规模数据的随机、实时读写访问，同时可以使用MapReduce来处理其保存的数据，它将数据存储和并行计算完美地结合在一起[10]。也就是说，HDFS为HBase提供了高可靠性的底层存储支持，MapReduce为HBase提供了高性能的计算能力。除了HBase，还有其他存储框架，比如ElasticSearch、Cassandra、Redis、MongDB等。MapReduce和HBase都是Hadoop生态系统的核心组件，各组件间密切结合的设计原理的一大优点是能够构建出无缝整合的不同处理模型的应用[11]。鉴于此，适合采用HBase存储模拟结果解析后的结构化数据（记录集）和非结构化数据（图片集）。大数据存储框架与机理模型融合的核心在于将分布于各个计算节点的模型计算原始结果文件和解析后的数据记录并发写入持久化存储设备。

针对模型模拟结果挖掘分析，机理模型与大数据融合体现在对模型结果的快速提取及挖掘分析。在数据分析方面，首先要求数据处理速度足够快，速度快意味着可以满足交互式查询的需求；其次，要求剥离对集群本身的关注，不需要关注如何在分布式系统上编程，也不需要过多关注网络通信和程序容错性，只需要专注于满足不同应用场景下的需求；最后，要求支持通用的交互式查询、机器学习、图计算等不同运算，且能通过一个统一的框架支持这些计算，从而简单、低耗地把各种处理流程整合在一起，这样的组合在实际的数据分析过程中很有意义，减轻了对各种分析平台分别管理的负担。Spark是满足上述需求的一个快速大数据分析框架。Spark于2009年诞生于加州大学伯克利分校RAD实验室，其一开始就是为交互式查询和迭代算法设计的，同时支持内存式存储和高效的容错机制。Spark支持在内存中进行计算，因而具有快速的处理速度，支持交互式查询。Spark包含多个紧密集成的组件，比如Core（任务调度、内存管理、错误恢复、存储交互）、SQL（操作结构化数据）、Streaming（实时计算）、MLib（机器学习算法库）、GraphX（图计算库）等，各组件间密切结合，支持各种各样的应用需求。和HBase一样，Spark也是Hadoop生态系统中的核心组件之一。鉴于此，适合采用Spark框架对模拟结果进行

进一步分析，典型应用功能包括交互式查询、模型参数敏感性分析、模型率定、模型验证、模型预测和应用决策。大数据分析框架与机理模型融合的核心在于快速提取分布于多个存储节点上的模拟结果，组织成物理上分散、逻辑上统一的结构化数据格式，依托已有算法库进行数据分析。

4 水环境模型与大数据技术融合的技术思路

下面分别从水环境模型的规模计算、规模存储和应用分析角度，阐述实现融合框架的技术思路，具体如图3所示。

4.1 水环境模型的规模计算

参考文献[9]最早采用Hadoop1.0开展水环境模型和大数据计算框架融合的研究，通过将SWAT模型率定和不确定性分析中的规模运算分解到map和reduce过程，为解决水文建模中的计算需求问题提供了一种有效方法。受此启发，参考文献[8]等提出一种仅使用map过程的改进方法，基于Hadoop2.0实现水动力水质模型Delft3D的集群运算架构。该架构不使用shuffle过程，提高了计算运行效率，为解决水环境模拟规模计算问题提供了新的视角。上述两者指明了在统一平台内耦合数值模型和计算框架的方法，即MapReduce模式下，模型作为第三方可执行程序被批量调用，map负责分布式计算，reduce负责汇总结果。

水环境模型具有其内在特点：一方面，包括流域分布式水文模型SWAT和三维水动力水质模型Delft3D在内的水环境数值模拟模型等，通常属于CPU密集型计算，运行

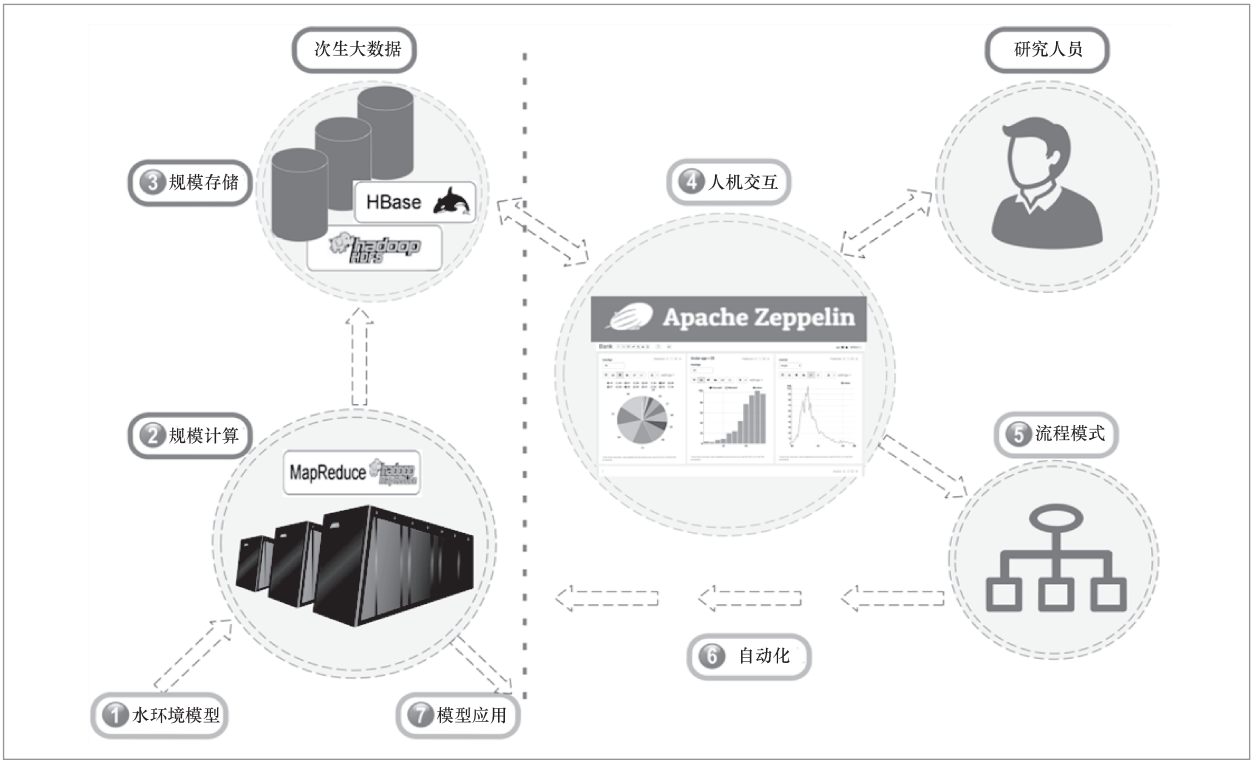

图 3 水环境模型与大数据技术融合的技术思路

时间范围为几分钟到几小时，甚至到几天。这和MapReduce的设计目标相符[12]；另一方面，成熟的机理模型通常基于服务于科学计算的FORTRAN语言编写，大数据计算框架往往基于Java语言编写，两套语言混合编程需要付出极大的开发成本，在实践中应该发挥各自语言的特色，控制或降低融合成本。综上所述，机理模型与大数据计算框架融合的思路是将机理模型作为独立的第三方可执行程序被计算框架调用，所有计算任务之间相互独立，不发生交互。

从实现角度，为了提高模型计算速度，通常需要保障独立的计算核和足够的内存。不论是Hadoop1.x中的JobTracker还是Hadoop2.x中的MRAppMaster都有整体计算资源的管理机制，都可根据应用需求动态调配内存和计算核。目前，YARN被证明是一个有效的资源管理工具，而且和MapReduce一样同属Hadoop生态圈，这方便了模型计算资源的调配。完成资源调配后，进入模型运行环节，这是模型和计算框架融合的核心，此环节主要包括模型启动、状态追踪以及管理交互（暂停、继续、中止和重启等）。典型的Hadoop集群通常运行在Linux操作系统下，因此，模型引擎首先需要在Linux操作系统下编译才能被使用；其次，模型在运行之前，需在集群中能够被获取；最后，在本地节点中模型被调用进而触发执行过程。

本质上，模型作为第三方程序，通常不能将执行状况告知节点，因此需要构建一套反馈机制以发送状态给用户。详细的状态信息应该包括两部分，即模型自身运行状态和任务运行状态，用户获取到这些信息后会决定是否干预运行状态。

综上所述，模型和大数据计算框架融合的核心在于发挥计算框架的优势，为模型的规模计算提供一种理想的方法。在具体的融合过程中，首先，依赖计算框架的资源调配机制，确保满足模型对计算资源的需求；其次，依赖计算框架的编程模型，确保批量计算数据被分发，计算程序在分布式节点上可获取、可调用和状态可追踪。

4.2 水环境模型模拟结果规模存储

数据的持久化是一个非常重要的过程，存储结构的设计决定了后续数据处理和分析的效率。当模型计算完成时，模拟结果存储过程随之启动。为了实现高效持久化，结果文件一般以二进制格式存储。不同的模型有不同的存储格式，比如Delft3D的NEFIS格式、MIKE的dfs2格式等，因此需要不同的解析方法，以便提取重点关注的信息并将其写入存储介质。一般来说，模型说明文档提供了相应的文件结构。根据此结构，可以设计合适的存储结构。文件和数据库是两种比较常见的存储模式。由于计算会耗费大量时间，原始结果文件显得弥足珍贵，采用分布式文件存储系统来存储原始结果文件成为一种合理选择。HDFS内置的可扩展、冗余、容错机制确保了原始数据的可靠性和高可用性，使其成为一种较佳的选择。其他选择还有使用同样广泛的FastDFS，具体采用哪种存储系统，需要结合计算框架和分析框架，从能否无缝结合的角度进行综合考虑。

通常情况下，原始结果文件不能被直接使用，实际应用中需要借助专业的后处理工具。结构化的表结构是广泛应用的一种形式。因此，通常需要将原始结果集进行解析和提取，形成结构化数据集。根据水环境管理领域特点，这种结构化数据集是一种典型的时间序列数据集。因此，在存储阶段，除了需要关注规模化的并行写入，还要兼顾数据本身的时间序列特点。分布式列式数据库在海量数据并行写入方面具有优势，但是考虑到数据本身的时

间特性，时间序列数据库是一个合适的选择，而且其存储框架和计算框架还能属于同一个生态系统。不幸的是，在Hadoop生态系统中，在存储方面除了扩展HBase的OpenTSDB，没有单独提供时间序列数据库。即使OpenTSDB提供对时间序列存储的支持，通过对HBase的存储模式和行键同时进行优化设计，也能够获得优于OpenTSDB的读写性能。此外，其他选择还有InfluxDB、RRDtool、Graphite和ElasticSearch等数据库。因此，对于模拟结果存储，有很多的选择。尽管如此，不论采用何种数据库，都需要慎重考虑存储结构设计和主键设计，二者共同决定了后文中应用分析的效率。

综上所述，模型和大数据存储框架融合的核心在于选择合适的模型计算结果持久化框架。在具体融合过程中，持久化存储需要兼顾原始结果和解析记录两种不同格式。首先，需要判断位于多个计算节点上的计算程序是否运行结束；其次，节点上的计算完成后，在本地节点上连接并写入存储框架，实现多节点并发存储。

4.3 水环境模型模拟结果应用分析

模型结果分析依赖两部分：结果集的获取、基于数据之上的算法以及交互模式。在上述模型与存储融合中，数据以分布式文件和数据库的形式存储，推荐采用的数据库分别是HDFS和HBase。对于HDFS，Hadoop提供了客户端FileSystem对象，该对象提供了类似操作本地文件的方法，可以将整个HDFS视作一块大硬盘来处理。借助FileSystem，可以将HDFS中存储的原始结果文件和解析后制图渲染出的批量制图文件下载到本地，其中，批量制图文件可以直接使用，原始文件则需要进一步读取和深入分析处理。对于分布式文件的处理，MapReduce通常被认为是一个最佳选择。MapReduce擅长批量的顺序处理，但是不支持随机查询。为此，需要一种支持随机查询的框架。对于HBase中存储的记录集，采用主流的Spark计算框架进行分析成为适宜的选择。利用其内置的HBase-Spark接口，可以在HBase中进行交互式数据检索，并将检索结果转化为易于操作的DataFrame格式。DataFrame派生于弹性分布式数据集（resilient distributed dataset，RDD），RDD是Spark SQL模块中最核心的编程抽象，可以被理解为以列的形式组织的分布式数据集合，它类似于关系型数据库中的表，但在底层实现优化并提供了一些抽象的操作来支持SQL。与此同时，Spark内部支持随机查询，这也为HDFS原始文件处理提供了支撑[13]。

RDD结构充分凸显了分布式数据处理的优势。RDD是逻辑集中的实体，在集群中的多台机器上进行了数据分区。通过对多台机器上不同RDD分区的控制，可以减少机器之间的数据重排。RDD是Spark的核心数据结构，具有丰富的算子以支持复杂分析，使用Spark集群的计算资源。在不影响HBase集群稳定性的情况下，可以通过并发分析的方式提高Spark的性能。RDD支持两种操作，转换操作用于将一个RDD转换生成另一个RDD，行动操作则触发Spark提交作业，并将数据输出到Spark系统。基于RDD构成了Spark的四大组件，SQL用于处理结构化数据，Streaming用于处理流数据，MLlib用于机器学习，GraphX用于图计算。因此，Spark提供了用于大规模数据处理的统一分析引擎，原始的结果文件和HBase中的解析结果集可以转换为Spark的RDD，充分利用Spark生态系统实现算法高效处理分析。

交互和可视化在知识发现中非常重

要。Spark生态系统中提供Shell和Submit两种客户端来支持交互，但是缺乏支持可视化分析的组件。对此，Hadoop生态系统中的Zeppelin组件同时被赋予交互和可视化能力，Zeppelin是一个可以进行大数据可视化分析的交互式开发系统，可以承担数据接入、数据发现、数据分析、数据可视化、数据协作等任务。Zeppelin前端提供丰富的可视化图形库，后端支持HBase、Flink等大数据系统，并支持Spark、Python、Java数据库连接（Java database connectivity，JDBC）、Markdown、Shell等常用解释器，这使得开发者和研究者可以方便地在Zeppelin环境中进行数据分析。Zeppelin原生支持Spark解释器，使得其成为一个合适的数据分析及可视化工具。

综上所述，模型和大数据分析框架融合的核心在于选择合适的交互式分析框架，该框架支持数据快速获取、高效组织、交互式和可视化分析。具体融合流程分为3步：首先，借助交互式可视化工具，探索隐藏在数据中的规律，形成专业应用功能；其次，梳理上述交互式探索的流程，采用程序语言将处理流程封装为算法软件包；最后将算法软件包集成到分布式计算框架中，用于服务具体业务的应用需求。

5 水环境模型与大数据技术融合案例

为了证实机理模型与大数据技术融合的可行性，以江西省梅江流域SWAT模型率定为例，描述水环境机理模型与大数据技术融合的具体实现流程。限于篇幅，本文仅给出融合过程中关键技术的实现流程。具体SWAT空间建模细节可参考参考文献[14]。

5.1 SWAT模型空间建模

SWAT是一种半分布式模型，已被广泛用于水文和环境建模。SWAT已开放源代码软件，其中的许多数据库、文档和出版物可供公众使用，因此备受关注。此外，SWAT还提供了几种软件产品（如ArcSWAT、QSWAT），可以提供友好的用户界面，并以直观、信息丰富的地图形式显示结果。如**图4**所示，SWAT的输入数据包括数字高程模型（digital elevation model，DEM）、河道、土地利用和土壤数据。

构建SWAT模型的流程如下：

- 将河道的线状几何信息加载到DEM；
- 对DEM进行预处理，并指定最小子流域面积；
- 核对并编辑河道上的点状要素，选择和定义流域出水口；
- 完成流域划分；
- 根据土地利用、土壤数据和坡度数据对研究区域进行分类，以确定每个子汇水区的水文响应单元（hydrological response unit，HRU）；
- 将所有天气数据作为时间序列表输入模型；
- 基于坡度、土地利用、土壤数据和天气条件估算模型参数；
- 对SWAT模型进行校准和验证，在模型运行过程中，最好设置一个预热期以避免初始化错误。

5.2 SWAT模型与大数据技术框架融合

与传统并行计算模式不同，基于大数据技术中的分布式并行计算框架实现的集群运算模式属于多算例多任务分解模式，即每一个算例对应一个SWAT模型运

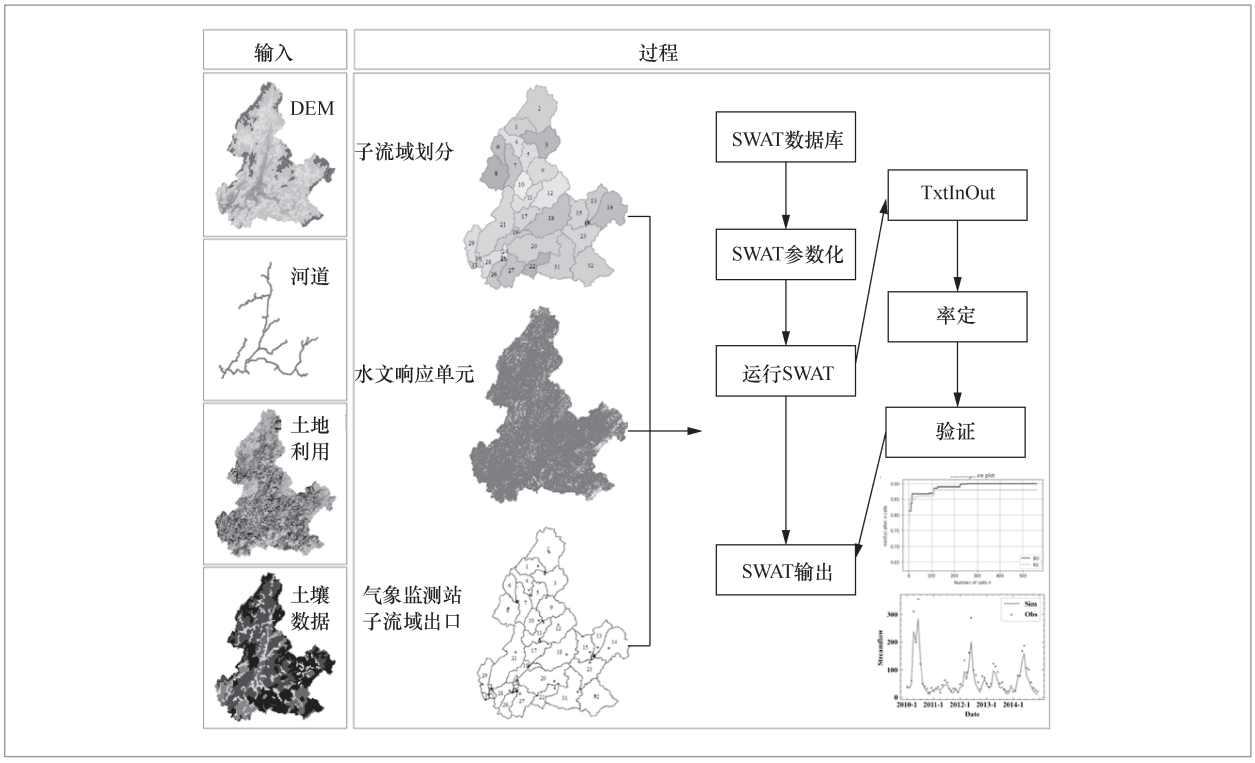

图 4 SWAT 模型空间建模流程

行，这种模式适合大批量模型计算。通过位置感知将计算移动到数据所在的存储位置是一个重大的进步[15]，即通过“数据本地化”可减少数据迁移，从而节约网络带宽，获得高效的计算性能。分布式存储将SWAT模拟结果分散存储到多个节点，并且同一份数据在不同节点上保存多个副本，兼顾实现数据本地化和冗余备份，保障了数据的安全性。分布式计算则通过位置感知将SWAT模型的可执行程序分发到案例配置文件所在位置，达到“计算本地化优化”的目标。

SWAT模型在Hadoop平台下的集群运算模式如图5所示。配置文件的分布式存储冗余备份机制缩短了计算程序的寻址感知时间。SWAT模型的分布式计算包括位置感知、本地化计算和计算结果存储3个过程。分布式分发机制可以快速定位配置文件所在的计算节点，自动下载SWAT模型执行文件到计算节点，并创建运行空间，启动模型读取配置文件，执行模型本地化计算，最后将SWAT模拟结果写入分布式存储。

5.3 案例研究：SWAT模型自动率定

本案例采用贝叶斯优化（Bayesian optimization，BO）算法对SWAT模型进行参数估值，其率定流程如图6所示。复杂水质模型包含大量参数，但通常只有少数参数会影响模型的输出[16]。鉴于对不重要或者不敏感的参数进行估值会导致模型的过度参数化，从而大大降低参数估值的效率[17-18]，在进行参数估值之前需要进行敏感参数选择。本案例采用Morris敏感性分析方法，对选择的重要参数进行敏感性排序，并筛选出敏感参数；然后，采用贝叶斯优化算法对筛选出的参数进行估值；最后，通过分析

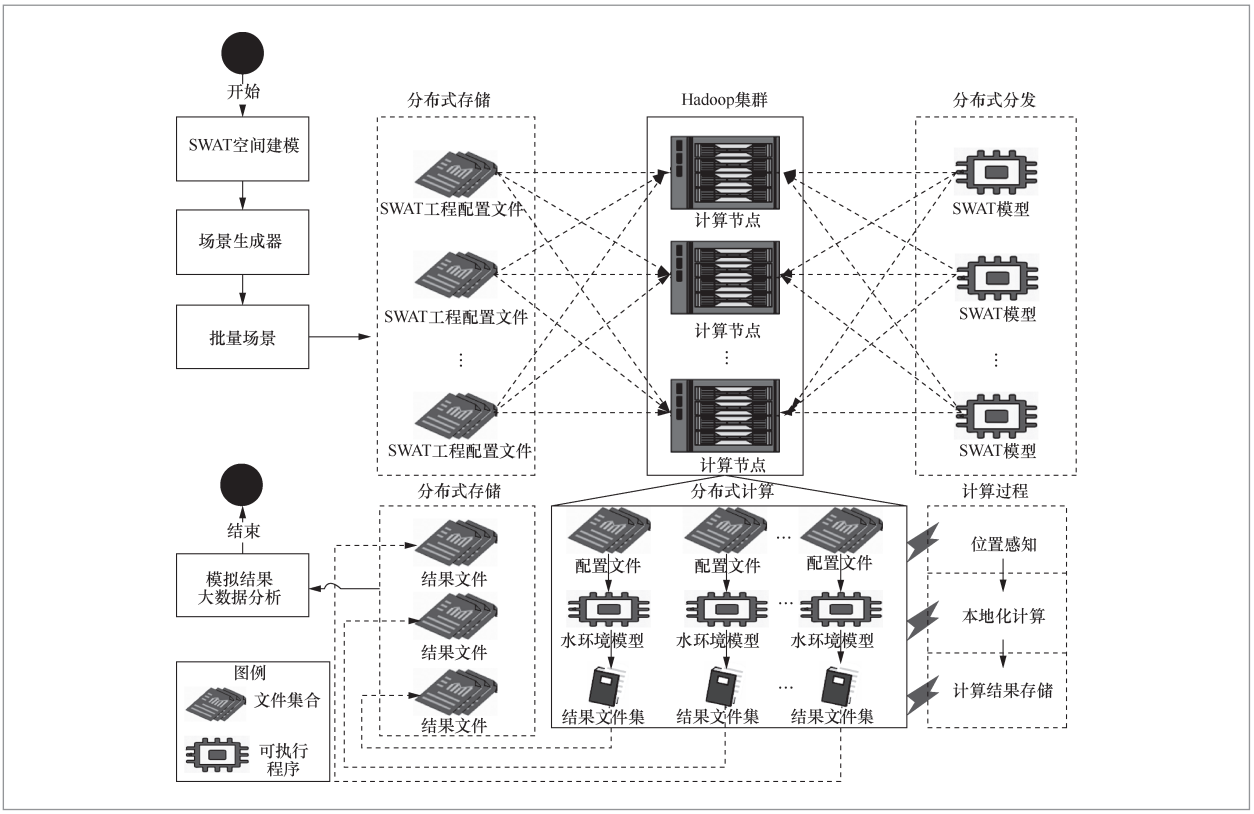

图 5 SWAT 模型在 Hadoop 平台下的集群运算模式

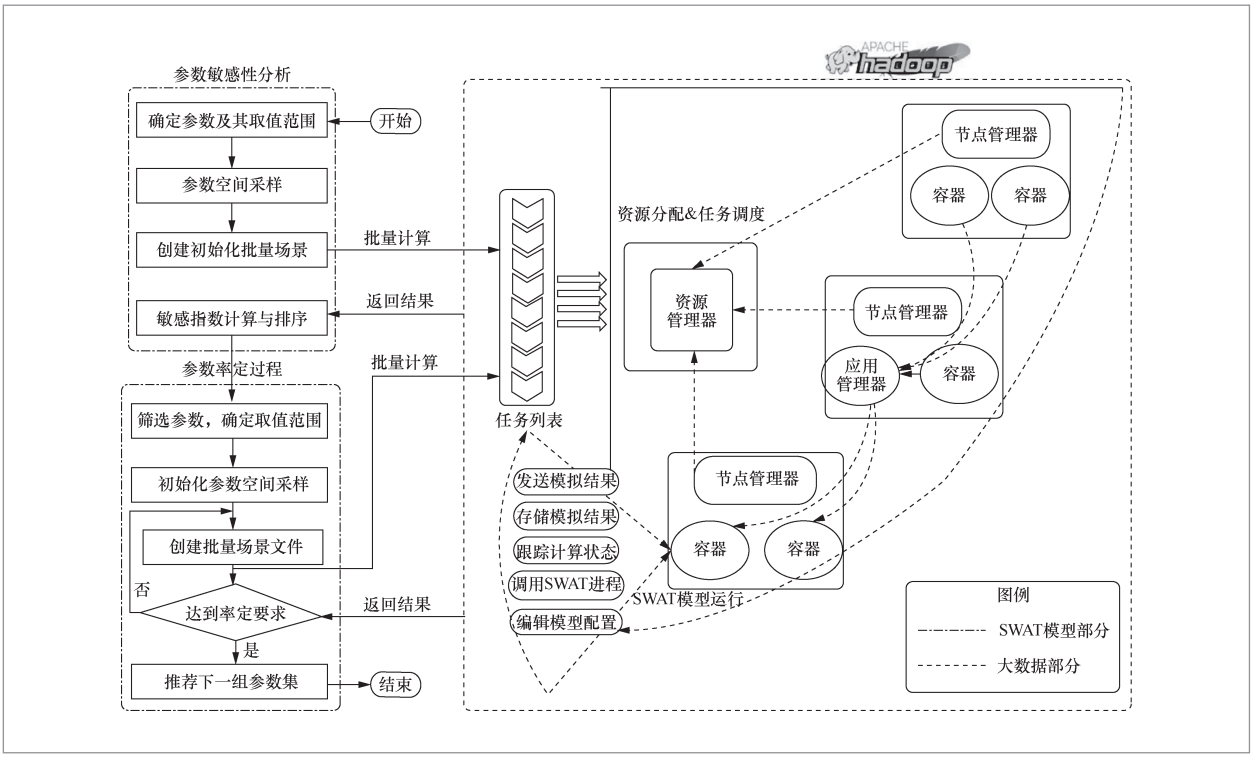

图 6 SWAT 模型自动率定流程

参数估值方法的优化效率和水质模型的拟合效果，对方法的适用性进行评估。

Morris敏感性分析法是由Morris于1991年提出，后经过改进的一种全局敏感性分析方法[19]。该方法适用于分析参数众多且运算量较大的模型，被广泛应用于因子固定（factor fixing）和敏感性分析中。该方法的优点是以较低的计算成本获得模型参数的敏感性相对大小，并对模型参数的敏感性大小进行排序[20]。基于高斯过程的贝叶斯优化算法具有收敛速度快、优化迭代次数少的特点，适用于解决评估代价高昂的环境模型的自动参数率定问题。纳什效率系数（Nash-Sutcliffe efficiency coefficient，NSE）被当作目标函数进行水文参数敏感性及收敛性分析，从而定性分析度量方式的影响。NSE是最常用的模型评价指标之一[21-22]，表示模型拟合方差占总方差的百分比[23]。

5.4 结果

SWAT模型参数选择及其范围是根据以往的研究确定的，选用参考文献[24]中推荐的27个参数，经Morris采样后产生24 000组参数组合。进行**图6**中的批量计算后，返回同等数量的NSE结果。将24 000组参数集-NSE结果数据输入Morris算法，得到敏感指数并排序。选择敏感指数大于0.08的8个参数作为率定参数。各参数名称、最大/最小值及参数意义见**表1**。值得注意的是，CH_K2和CH_N2属于可测量参数，由于在本例中并没有测定，因此仍然将其作为率定参数。

本案例中的大数据集群软硬件信息详见参考文献[8]中的**表1**，贝叶斯优化算法的应用流程为：初始化参数空间，进行拉丁超立方抽样，据此创建批量模型场景文件；将场景文件分发到大数据集群，集群在文件所在节点执行模型计算，返回NSE目标函数值；根据NSE目标函数值判断是否达到率定要求，当未满足率定要求时，选择下一组参数继续进行迭代计算，反之终止集群运算，返回率定结果。详细操作流程可参考参考文献[25]。

考虑到贝叶斯优化算法本质上属于概率优化，因此本案例采用20次重复测试以检验优化结果的合理性。结合大数据集群硬件性能指标（56个算例并发计算），在单次优化过程中设置560（56×10）次迭代计算。**图7（a）**显示了某次迭代过程中，NSE随着迭代次数的增加而缓慢增长的趋势。结果表明，NSE总体上保持逐步上升趋势。**图7（a）**也有助于帮助理解贝叶斯优化过程，即如何推荐下一组参数集取决于开发和探索之间的权衡，NSE逐步上升过程中存在的波动变化证明了这一点。

为了进一步证明贝叶斯优化算法在SWAT参数率定中的优势，本案例使用随

表1 SWAT模型中识别的敏感性参数

参数名称	最小值	最大值	参数意义
CN2	35	98	径流曲线数
ALPHA_BNK	0	1	河岸基流α因子
SOL_K	0	2 000	饱和导水率
CH_K2	-0.01	500	主通道冲积层中的有效导水率
CANMX	0	100	最大冠层储藏量
SOL_AWC	0	1	土壤有效含水率
SOL_BD	0.9	2.5	土壤密度
CH_N2	0	0.3	主河道曼宁系数

机搜索（random search，RO）算法进行对比。从**图7（b）**可以看出，当BO和RS达到同一较高NSE取值时（以0.87为例），BO迭代次数（121）略小于RS（130），说明BO率定效率高于RS。而在给定迭代次数下（以220为例），BO优化获得的NSE（0.89）大于RS（0.88），说明BO率定效果优于RS。上述结果表明，从SWAT参数率定和优化的效果和效率角度来看，贝叶斯优化算法优于随机搜索算法。此外，依托大数据技术框架，可以融合多种优化算法，建立水环境模拟优化框架，从而为水环境模型深度应用和价值发挥提供理想环境。

为了检验本案例的率定效果，对径流模拟值（Sim）与实测值（Obs）进行对比，如**图8（a）**所示。BO的NSE最大值为0.89，说明率定后的模型较好地捕捉了月径流变化，可以用于案例月径流模拟。对贝叶斯优化过程中产生的20×560个径流模拟结果值进行概率统计，获得径流模拟结果的概率直方图，如**图8（b）**所示。从**图8（b）**可以看出，在率定过程中，径流模拟值近似呈正态分布状收敛，概率分布中极大值处的径流量趋近于观测值，说明率定过程中的参数不确定性可以传到模拟结果中，从而导致模拟结果的不确定性。因此，在实际应用中，模拟结果分布呈现出的收敛趋势

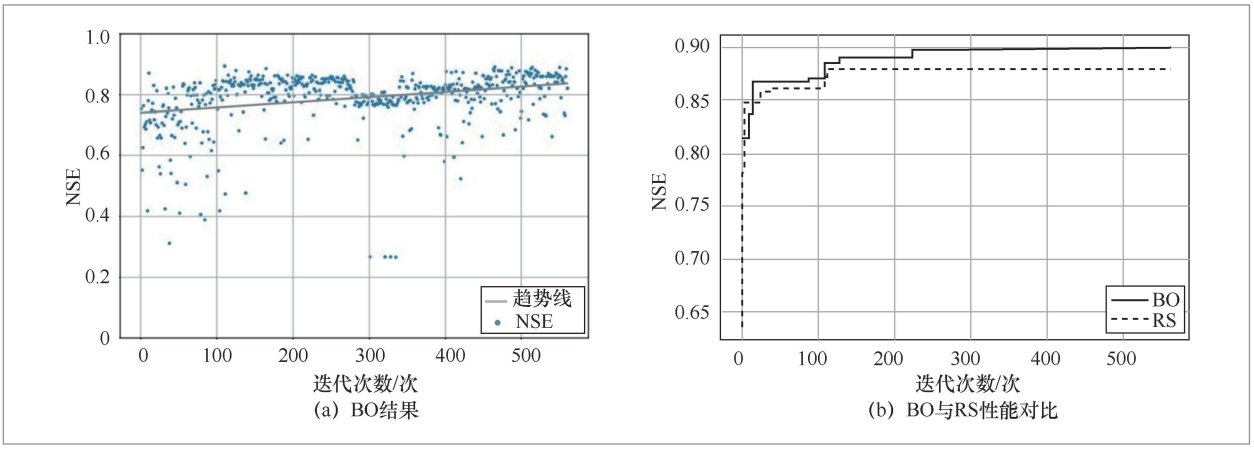

图 7 案例率定效果验证一

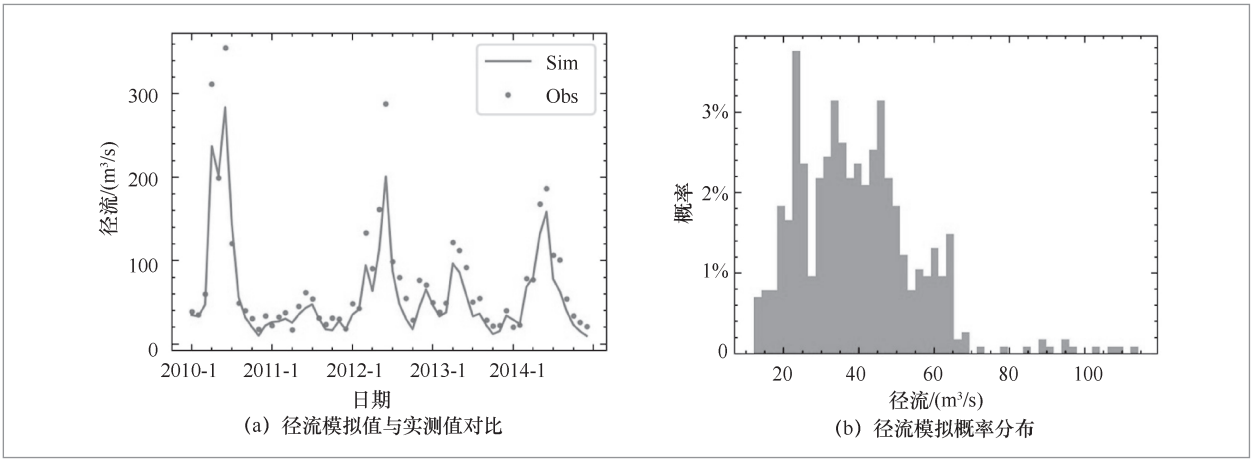

图 8 案例率定效果验证二

比取某个特定值更具有参考价值，同时也反映了开展模型不确定性研究的重要性，本案例中提出的模型与大数据技术融合策略可以为开展模型不确定性研究提供一种新思路。

6 结束语

限制水环境模型成功应用的主要瓶颈是基础数据难以获取以及模型率定、模型验证及场景分析中的高负荷计算。基础数据获取依赖于精准的长期监测和对监测信息的高效提取，也可采用理论和经验相结合的方法弥补数据的欠缺。大数据存储技术的可扩展、冗余、容错机制确保了原始数据的可靠性和高可用性，使其成为一种合适的持久化选择，为多源异构基础数据的持久化存储提供了解决方案。作为典型的计算密集型复杂系统模型，水环境模型通常需要大量计算时间，尤其在面向自动率定、验证及场景分析等批量计算需求时，通常无法承受大量的迭代计算。在单个模型计算非常耗时的情况下，批量计算是被禁止的。虽然现有并行计算体系很好地解决了数值模型的高性能计算问题，但是在计算结果的规模存储，尤其规模分析上性能表现一般。这和现有并行计算体系的设计目标有关，它注重计算的高效性，而未考虑其他需求（如存储和分析需求）。因此，水环境模型的高质量应用迫切需要一个紧密衔接计算、存储和分析全链条的技术支撑体系。大数据技术体系成为潜在的理想选择。大数据技术内置了分布式计算、存储和分析框架体系，自然成为一种解决水环境模型规模计算问题的潜在理想方案。

在本研究中，以SWAT模型参数自动率定为例，验证了方案的可行性。首先，通过将模型配置文件分布式分发到各个计算节点；接着，水环境模型计算程序会自动定位到配置文件所在位置；然后，在计算节点启动计算的过程中，在SWAT模型计算完成后，开始解析计算结果，并将原始结果文件和解析结果记录存入大数据库；最后，利用内存网格技术高效地提取和分析模拟结果。上述所有环节紧密衔接了计算、存储和分析技术链条，应用案例证明了水环境模型与大数据技术融合的可行性，二者的融合为深入挖掘水环境模型的应用潜力和充分发挥其应用价值提供了新的视角。

水环境模型和大数据技术融合的核心是模型分布式计算，即模型作为独立的第三方可执行程序被计算框架调用，比较适合模型参数率定及情景分析等批量计算的应用场景。受益于大数据分布式计算横向扩展的特点，计算效率通常和计算节点个数呈线性增长关系，这极大提高了模型的计算效率。即便如此，作为计算密集型复杂模型，水环境模型计算仍然非常耗时。相反，近似物理模型的统计“代理模型”可以提供对物理系统的高效仿真。代理模型系统以统计模型的形式映射输入变量和输出变量，该统计模型通过使用物理模型生成的一组数据进行训练和验证。代理模型在水文学领域已被广泛研究[26]，近似算法（如kriging[27]、人工神经网络[28]、径向基函数[29]、多项式回归[30]、支持向量机[31]、稀疏网格插值法[32]和随机森林技术[33]）已被应用于各种地球系统和水文系统。在最新研究中，复杂水动力水质模型EFDC（environmental fluid dynamics code）被长短期记忆（long short-term memory，LSTM）代理反映了这一趋势[34]。可以预见，水环境模拟与大数据技术融合有以下两个发展趋势。

（1）以大数据技术为转化载体，水环境模型将以统计模型形式从物理模型转化为代理模型，这将极大地改变现有水环境

模型的应用模式。与物理模型相比，代理模型兼具较高的模拟精度和极高的计算效率，使之成为物理模型的理想替代，这势必会推动模型参数敏感性分析、参数率定及情景分析等应用研究。

（2）水环境模拟优化框架成为未来的发展趋势。该框架以完整的分布式计算、存储和分析链为技术支撑，以物理模型或代理模型为核心，结合单目标或多目标优化算法，解决优化调控类科学决策问题。

参考文献:

[1] National Academies of Sciences, Engineering, and Medicine, Division on Earth and Life Studies, Water Science and Technology Board, et al. Future water priorities for the nation: directions for the U.S. geological survey water mission area[M]. Washington D.C.: National Academies Press, 2018.

[2] REICHSTEIN M, CAMPS-VALLS G, STEVENS B, et al. Deep learning and process understanding for data-driven earth system science[J]. Nature, 2019, 566(7743): 195-204.

[3] 陈永灿, 刘昭伟, 朱德军. 水动力及水环境模拟方法与应用[M]. 北京: 科学出版社, 2012.
CHEN Y C, LIU Z W, ZHU D J. Hydrodynamic and water environment simulation methods and applications[M]. Beijing: Science Press, 2012.

[4] 季振刚. 水动力学和水质——河流、湖泊及河口数值模拟[M]. 李建平, 冯立成, 赵万星, 译. 北京: 海洋出版社, 2017.
JI Z G. Hydrodynamics and water quality: modeling rivers, lakes, and estuaries[M]. Translated by LI J P, FENG L C, ZHAO W X. Beijing: China Ocean Press, 2017.

[5] RAZAVI S, TOLSON B A, MATOTT L S, et al. Reducing the computational cost of automatic calibration through model preemption[J]. Water Resources Research, 2010, 46(11).

[6] ZHANG X S, SRINIVASAN R, BOSCH D. Calibration and uncertainty analysis of the SWAT model using genetic algorithms and bayesian model averaging[J]. Journal of Hydrology, 2009, 374(3-4): 307-317.

[7] LIN F R, WU N J, TU C H, et al. Automatic calibration of an unsteady river flow model by using dynamically dimensioned search algorithm[J]. Mathematical Problems in Engineering, 2017(7): 1-19.

[8] 马金锋, 唐力, 饶凯锋, 等. Hadoop下水环境模拟集群运算模式[J]. 大数据, 2019, 5(6): 73-84.
MA J F, TANG L, RAO K F, et al. Cluster computing mode for water environment simulation based on Hadoop[J]. Big Data Research, 2019, 5(6): 73-84.

[9] ZHANG D J, CHEN X W, YAO H X, et al. Moving SWAT model calibration and uncertainty analysis to an enterprise Hadoop-based cloud[J]. Environmental Modelling & Software, 2016, 84: 140-148.

[10] GEORGE L. HBase: the definitive guide(fisrt edition)[M]. Sebastopol: O'Reilly Media, 2011.

[11] KARAU H, KONWINSKI A, WENDELL P, et al. Spark快速大数据分析[M]. 王道远, 译. 北京: 人民邮电出版社, 2015.
KARAU H, KONWINSKI A, WENDELL P, et al. Learning Spark: lightning-fast data analysis[M]. Translated by WANG D Y. Beijing: Posts & Telecom Press, 2015.

[12] WHITE T. Hadoop: the definitive guide(second edition)[M]. Sebastopol: O'Reilly Media, 2010.

[13] 吴信东, 嵇圣硙. MapReduce与Spark用于大数据分析之比较[J]. 软件学报, 2018, 29(6): 1770-1791.
WU X D, JI S W. Comparative study on MapReduce and Spark for big data analytics[J]. Journal of Software, 2018, 29(6): 1770-1791.

[14] SHIVHARE N, DIKSHIT P K S, DWIVEDI S B. A comparison of SWAT model calibration techniques for hydrological modeling in the Ganga river watershed[J]. Engineering, 2018, 4(5): 643-652.

[15] MURTHY A C, VAVILAPALLI V K. Apache Hadoop YARN: moving beyond MapReduce and batch processing with Apache Hadoop 2[M]. Boston: Addison-Wesley Professional, 2014.

[16] YI X, ZOU R, GUO H C. Global sensitivity

analysis of a three-dimensional nutrients-algae dynamic model for a large shallow lake[J]. Ecological Modelling, 2016, 327: 74-84.

[17] SONG X M, ZHANG J Y, ZHAN C S, et al. Global sensitivity analysis in hydrological modeling: review of concepts, methods, theoretical framework, and applications[J]. Journal of Hydrology, 2015, 523: 739-757.

[18] JIANG L, LI Y P, ZHAO X, et al. Parameter uncertainty and sensitivity analysis of water quality model in Lake Taihu, China[J]. Ecological Modelling, 2018, 375: 1-12.

[19] CAMPOLONGO F, SALTELLI A. Sensitivity analysis of an environmental model: an application of different analysis methods[J]. Reliability Engineering & System Safety, 1997, 57(1): 49-69.

[20] 刘松, 佘敦先, 张利平, 等. 基于Morris和Sobol的水文模型参数敏感性分析[J]. 长江流域资源与环境, 2019, 28(6): 1296-1303.
LIU S, SHE D X, ZHANG L P, et al. Global sensitivity analysis of hydrological model parameters based on Morris and Sobol methods[J]. Resources and Environment in the Yangtze Basin, 2019, 28(6): 1296-1303.

[21] BENNETT N D, CROKE B F W, GUARISO G, et al. Characterising performance of environmental models[J]. Environmental Modelling & Software, 2013, 40: 1-20.

[22] BAE S, SEO D. Analysis and modeling of algal blooms in the Nakdong River, Korea[J]. Ecological Modelling, 2018, 372: 53-63.

[23] NASH J E, SUTCLIFFE J V. River flow forecasting through conceptual models part I—a discussion of principles[J]. Journal of Hydrology, 1970, 10(3): 282-290.

[24] GHAITH M, LI Z. Propagation of parameter uncertainty in SWAT: a probabilistic forecasting method based on polynomial chaos expansion and machine learning[J]. Journal of Hydrology, 2020, 586: 124854.

[25] 任婷玉, 梁中耀, 刘永, 等. 基于贝叶斯优化的三维水动力-水质模型参数估值方法[J]. 环境科学学报, 2019, 39(6): 2024-2032.
REN T Y, LIANG Z Y, LIU Y, et al. The parameters estimation method based on Bayesian optimization for complex water quality models[J]. Acta Scientiae Circumstantiae, 2019, 39(6): 2024-2032.

[26] RAZAVI S, TOLSON B A, BURN D H. Review of surrogate modeling in water resources[J]. Water Resources Research, 2012, 48(7).

[27] SIMPSON T W, MAUERY T M, KORTE J J, et al. Kriging models for global approximation in simulation-based multidisciplinary design optimization[J]. AIAA Journal, 2001, 39(12): 2233-2241.

[28] DOWLA F U, ROGERS L L. Solving problems in environmental engineering and geosciences with artificial neural networks[M]. Cambridge: MIT Press, 1995.

[29] REGIS R G, SHOEMAKER C A. A stochastic radial basis function method for the global optimization of expensive functions[J]. INFORMS Journal on Computing, 2007, 19(4): 497-509.

[30] FEN C S, CHAN C C, CHENG H C. Assessing a response surface-based optimization approach for soil vapor extraction system design[J]. Journal of Water Resources Planning and Management, 2009, 135(3): 198-207.

[31] ZHANG X S, SRINIVASAN R, LIEW M V. Approximating SWAT model using artificial neural network and support vector machine[J]. JAWRA Journal of the American Water Resources Association, 2009, 45(2): 460-474.

[32] ZENG L Z, SHI L S, ZHANG D X, et al. A sparse grid based Bayesian method for contaminant source identification[J]. Advances in Water Resources, 2012, 37: 1-9.

[33] BOYLE J S, KLEIN S A, LUCAS D D, et al. The parametric sensitivity of CAM5's MJO[J]. Journal of Geophysical Research: Atmospheres, 2015, 120(4): 1424-1444.

[34] LIANG Z Y, ZOU R, CHEN X, et al. Simulate the forecast capacity of a complicated water quality model using the long short-term memory approach[J]. Journal of Hydrology, 2020, 581: 124432.

作者简介

马金锋（1978– ），男，中国科学院生态环境研究中心助理研究员，主要研究方向为水环境数值模拟。

饶凯锋（1976– ），男，中国科学院生态环境研究中心助理研究员，主要研究方向为水生态毒理学、环境预警监测与物联网。

李若男（1982– ），女，中国科学院生态环境研究中心副研究员，主要研究方向为流域生态系统过程与模拟。

张京（1996– ），女，中国科学院生态环境研究中心硕士生，主要研究方向为水环境数值模拟。

郑华（1974– ），男，中国科学院生态环境研究中心研究员，主要研究方向为生态系统过程与生态系统服务。

收稿日期： 2021-02-03

通信作者： 郑华，zhenghua@rcees.ac.cn

基金项目： 国家重点研发计划资助项目（No.2019YFD0901105）

Foundation Item: The National Key Research and Development Program of China (No.2019YFD0901105)